链滴

SEO 实践（1）：开展 SEO 前的数据准备

作者：fc13240

原文链接：https://ld246.com/article/1600672330490

来源网站：链滴

当我们开始开展一项 SEO 工作时，第一件要做的事情是要保证我们做的任何事情都可以有数据支撑——而不是自己的直觉。SEO 的主要数据来源来自两块：网站的服务器日志、第三方流量分析具。
网站服务器日志
Apache，Nginx 等常用服务器的内置日志配置格式 Combine 已经可以满足大多数 SEO 分析需。它看上去类似是这样的：
111.111.111.111 – - "[20/Feb/2012:18:09:25 +0800]""GET / HTTP/1.1" 200 3121"http://***org/" "Mozilla/5.0 （compatible; Googlebot/2.1; +http://ww.google.com/bot.html）"
必须记录的信息诸如：访问来源 IP、访问时间、访问页面、HTTP 响应状态码、访问来源及客户标识等，这些在 Combine 日志格式里面都有。
在确保服务器日志可以满足其他部门的分析需求下，至少要确保上面提到的几项被记录在服务器志里面。但也不要将任何可以记录的数据都记录下来，只选择实际需要的部分，不然会使得网站日志积非常大，不利于分析起来的效率。这些内容可能需要和运维进行沟通解决。
然后关于日志的分析，我认为没太多固定的准备工作可做，因为它的数据来源是原始的（raw 似听上去会更有感觉？），所以可选择的数据维度几乎是无限的。因此尤其要按实际需求进行相应的处与分析。
对于一些要求并不是特别高的日志分析需求，可以尝试使用光年日志分析系统。虽然我个人对所图形界面的实用类程序都不带好感，但它提供了一些很不错的数据维度的思路。
听说有一家大型的旅游网站是采用 MongoDB 结合 Map/Reduce 进行日志分析的，我个人也用 MongoDB 实现过前面提到的光年日志分析的一部分重要功能。所以感觉 MongoDB 是个可以考虑选择。
第三方流量分析工具
Google Analytics 的安装
对于免费流量分析工具，Google Analytics 绝对是其中的佼佼者（以下简称 GA）。不过如果网的月浏览量大于 500W 的话，只有 Google Adwords 的用户，才能继续免费使用 GA 进行流量的记与分析。下面都以它为例。
在 GA 添加需要追踪流量的网站以后，它会提示你添加一段 JavaScript 代码，到每一个你需要踪页面的标记之前。代码的添加可能是一件很轻松的工作，但也可能非常麻烦，主要取决于网站的模层。
先提下常见开源博客程序 WordPress 的方法，它采用了包含的模板处理方式，比如网站首页、表页、文章页等自身的模板，都是只有当中一部分的。而包含网页 LOGO 等的网页头部，都使用 WodPress 的 get_header 方法来加载另一个独立的模板文件（get_header 方法本质上是 PHP 里面的 include 函数）。简言之，只要在 header.php 那个文件上面添加代码，包含它的所有网页都会跟着改很快就可以把 GA 代码添加好。
但情况并不总是理想的，尤其对于使用网站框架自己进行开发的网站，有时并没有将包含这样的式很好的运用。这可能是网站的建设规范不完善的关系，也可能是网站需求导致了确实无法使用和 WodPress 类似的包含方式。那么，至少要在每个网页的头部，额外包含一小段加载全局 JavaScript 的块，以方便的添加全局性的 JavaScript 代码。
虽然未必在添加 GA 代码时，对可能糟糕的网站模板结构去进行更改，最多到几十个不同的模板件里面去分别加下代码就是了（当然也要花些时间去保证没有漏过哪些页面）。但一次性搞定一些本性的问题会带来很多日后的便利性——比如又要换一套统计代码。
相对最麻烦的事情或许是如何说服程序员为了一些看似小的需求而修改模板结构，这边就略过了
一些基础的 Google Analytics 设置
对于 SEO 而言，一项最基础的设置，就是要把网站上对 SEO 有价值的页面进行归类。对页面进区分，并以此掌握了它们的流量现状及趋势以后，才能把握 SEO 的侧重点，及更好的分析网站上每次 SEO 修改的成效等等。
如最简单的例子，对于一个网站，如果手头有 1000 条外链，应该给网站的栏目页还是产品页？主要取决于哪类页面有更高的转化率与更大的 SEO 流量提升空间。

对于每个网站而言，都存在不同的情况。比如一个书籍类的电商网站，它列表页不会有太多流量没多少人搜索什么“计算机书籍”，但会更多人搜索《乔布什自传》之类，因为用户有很明确的需求而对于一个服饰电商，相应更多人会搜索“衬衫”之类，而非“2012 年春季新款白色衬衫”等，因用户只是想到网站上挑衣服，他们只有需求的意向，但具体需求是模糊的。
以上两个是比较典型的例子，但有更多情况我们无法用自己的直觉做出准确的判断，那就需要用量数据来收集事实。
尽管博客的流量数据分析起来没太大价值，出色的文章是博客的一切，但这里还是以 SEMWATC 为例来简单介绍下方法。假设我们需要把网站的栏目页和文章页流量进行区分，它们的 URL 分别是似这样的：/category/seo/，/2012/02/post/
首先要到 GA 的数据页面内，找到高级细分一项，点击右侧新自定义细分。然后进行类似下图的置：
通常情况下，将页面的 URL 匹配相应的正则以后，就可以把它们区分开来。注意，如果网站的期 URL 规划不完善，可能会导致无法用 URL 来区分页面类型的非常非常糟糕的情况，务必保证每一页面拥有其独立的 URL 标识。
在该例中，SEMWATCH 的栏目页匹配正则表达式是：^/category/.*?/，文章页是：^/2[0-9]{3}/[0-9]{2}/.*?/
尽量用最严格的正则表达式写法，这样可能可以在无形中规避很多不必要的错乱。还需要注意的，老版本的 GA 默认情况下筛选器的“包含”即使用正则表达式，新版 GA 一定要选择“匹配正则表式”这项。
关于正则表达式，篇幅所限不可能进行解释，如果你不懂的话，可以考虑去寻找程序员求助。但的个人建议是尽可能的要自己掌握它，这是一个比较基础的技术要求，SEO 不应该被它所难倒。正则达式虽然看上去很恶心——至少我从来看不懂自己写出来的正则，但其实挺容易学的。
总之通过上面的步骤，我们就简单的把页面类型区分开来了。回到最初的例子，如果有 1000 外给 SEMWATCH 随便分配，现在应该把外链给予哪些页面呢？可以发现的是栏目页几乎没流量、而章页天生流量就很高。多数情况下这证明了文章页具有更大的流量发展空间，此时把外链分配给文章就是最明智的做法。（但也不能武断的说，不能排除栏目页的 SEO 有巨大问题的可能性，这问题一都不罕见。所以还要结合我们的常识及其他方面的分析来综合判断。）
最后的总结
实际可能要面临的问题还有很多很多，当然不可能是一篇文章所能涵盖的。前面提到的只是两个要数据，实际 SEO 过程中，还或许需要用到的数据如网站级的 Google Webmaster Tool，估算流的爱站、SEMRush、Google Adplanner、HitWise，关键词的 Google Keyword Tool、百度司南链接类的 MajesticSEO、Ahrefs 等等。
最近我在看《麦肯锡方法》，提到：“以事实为基础，严格的结构化，以假设为导向”，类似的总结下 SEO 的话：“以数据为基础，严格的逻辑化，以效果为目标、技术为手段”。本文是为了作根基的数据垫下基础而已，它本身是没任何价值的——光看数据的话，它只不过是死板的数字罢了。
如何借由数据的辅助，在最需要的地方进行 SEO 的更改，使得流量获得大的突破并给网站产生值，这是我们要真正关注的部分，之后再慢慢分解。